Computational Social Science: Methods and Applications

Anjalie Field, anjalief@cs.cmu.edu

Overview

- Defining computational social science
  - Sample problems

- Common Methodology (Topic Models)
  - LDA
  - Evaluation
  - Limitations
  - Extensions

Definitions and Examples

What is Computational Social Science?

"The study of social phenomena using digitized information and computational and statistical methods" [Wallach 2018]

Social Science

- When and why do senators deviate from party ideologies?
- Analyze the impact of gender and race on the U.S. hiring system
- Examine to what extent recommendations affect shopping patterns vs. other factors

Explanation

Traditional NLP

- How many senators will vote for a proposed bill?
- Predict which candidates will be hired based on their resumes
- Recommend related products to Amazon shoppers

Prediction

[Wallach 2018]

How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not engaged argument [King et al. 2017]

- In 2014 email archive was leaked from the Internet Propaganda Office of Zhanggong

- Reveal the work of “50c party members”: people who are paid by the Chinese government to post pro-government posts on social media

Sample Research Questions [King et al. 2017]

- **When** are 50c posts most prevalent?
- What is the **content** of 50c posts?
- What does this reveal about overall government strategies?
-
- Additionally:
  - Who are 50c party members?
  - How common are 50c posts?

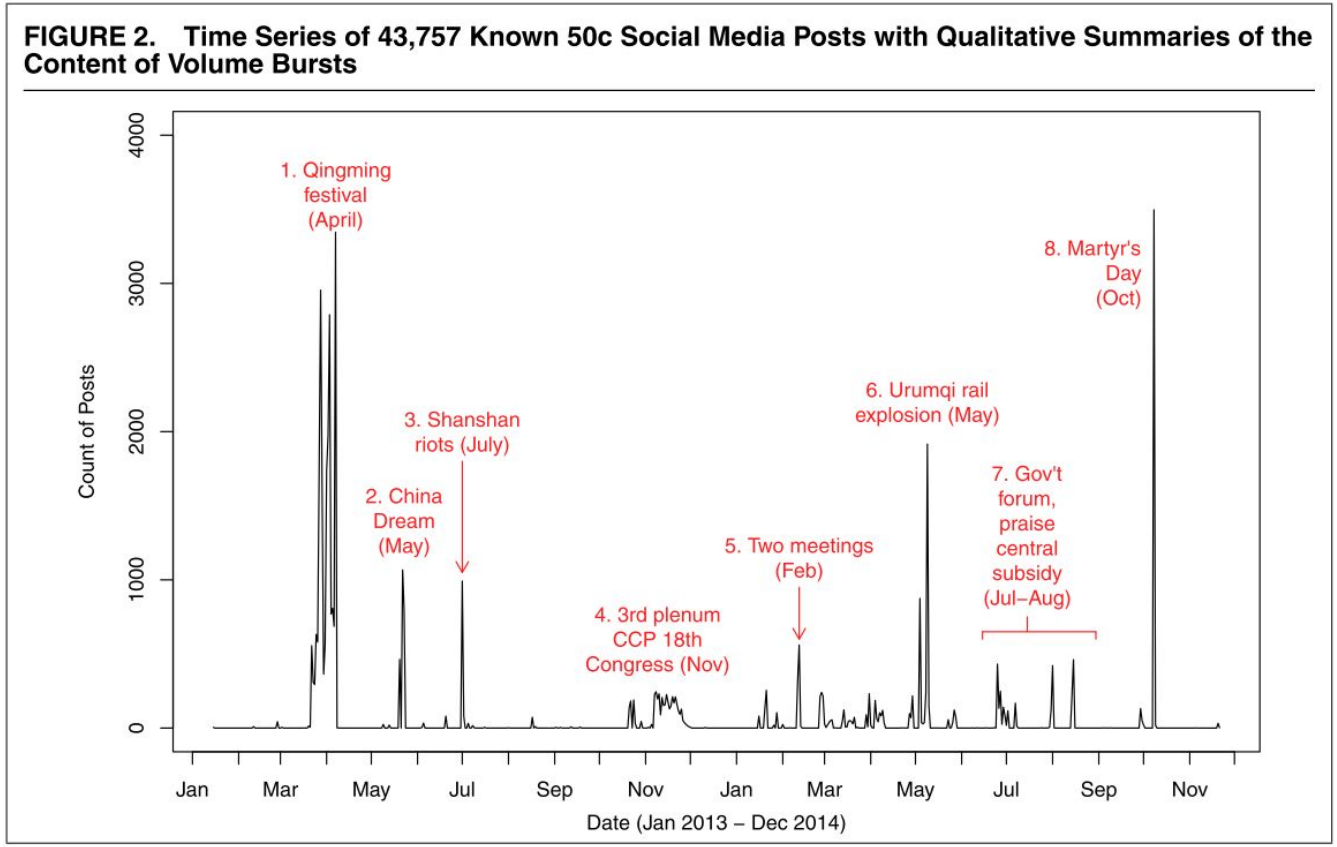
Preparations [King et al. 2017]

- Thorough analysis of journalist, academic, social media perceptions of 50c party members

- Data Processing
  - Messy data, attachments, PDFs

Preliminary Analysis [King et al. 2017]

- Network structure
- Time series analysis: posts occur in bursts around specific events

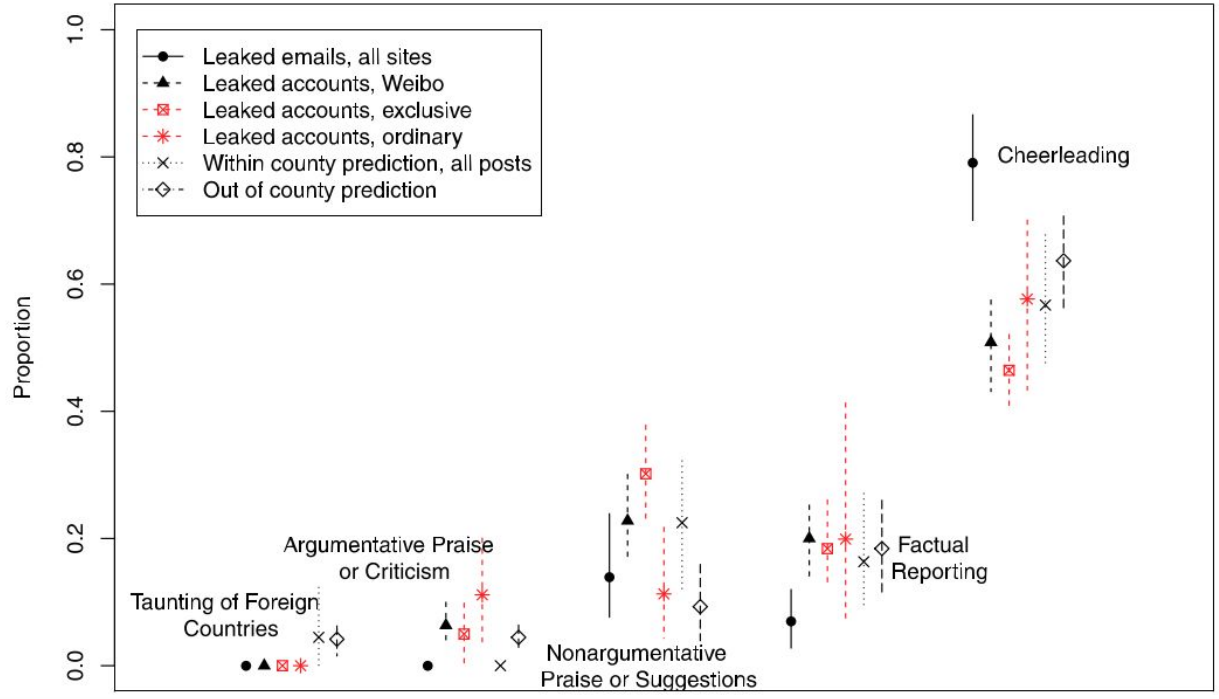


Content Analysis [King et al. 2017]

- Hand-code ~200 samples into content categories
  - Cheerleading, Argumentative, Non-argumentative, Factual Reporting, Taunting Foreign Countries
  - Coding scheme is motivated by literature review
  - Use these annotations to estimate category proportions across full data set
- Expand data set
  - Look for accounts that match properties of leaked accounts
  - Repeat analyses with these accounts
  - *Conduct surveys of suspected 50c party members*

Content Analysis [King et al. 2017]

FIGURE 3. Content of Leaked and Inferred 50c Posts, by substantive category (with details in Appendix A) and analysis (given in the legend)



Cheerleading: Patriotism, encouragement and motivation, inspirational quotes and slogans

Social Science

- Defining the research question is half the battle
- Data can be messy and unstructured
- Careful experimental setup means controlling confounds -- make sure you are measure the correct value
- Prioritize interpretability (plurality of methods)

Traditional NLP

- Well-defined tasks
- Often using well-constructed data sets
- Careful experimental setup means constructing a good test set -- usually sufficient get good results on the test set
- Prioritize high performing models

Twitter recently released troll accounts

- Information from 3,841 accounts believed to be connected to the Russian Internet Research Agency, and 770 accounts believed to originate in Iran
- 2009 - 2018
- All public, nondeleted Tweets and media (e.g., images and videos) from accounts we believe are connected to state-backed information operations

- **What can we do with this data?**

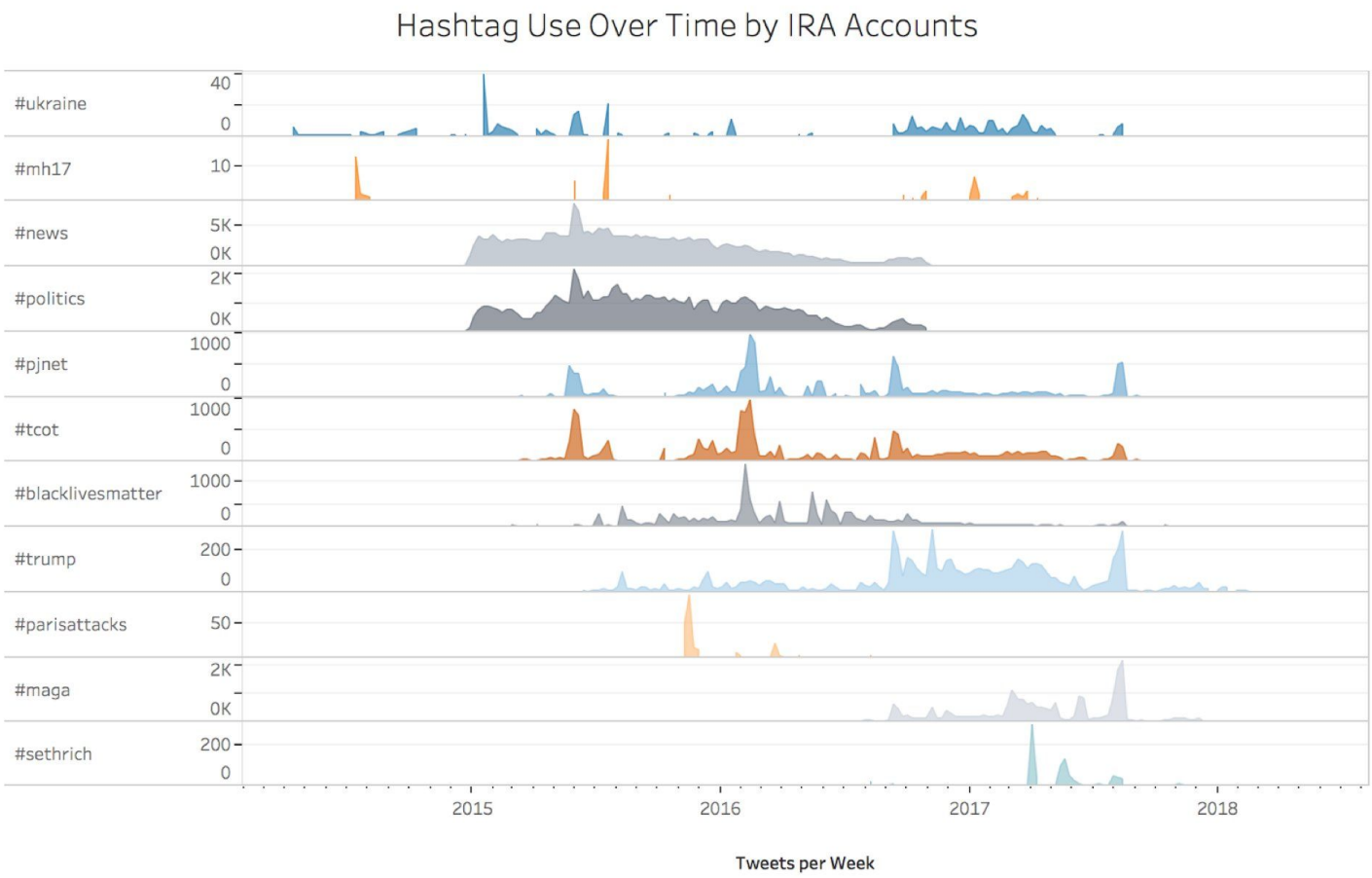
https://about.twitter.com/en_us/values/elections-integrity.html#data

What can we do with this data?

- *When* are posts most common? What events trigger tweets?
- What *content* is common? Argumentative? Cheerleading?
- What *stance* do tweets take? Do they take stances at all?
- What *impact* to tweets have? Which ones get favorited the most? Who follows/favorites them?
- *Who* do the tweets target? Who do the accounts follow?
- How much *coordination* is there? Do different IRA accounts retweet each other?

https://about.twitter.com/en_us/values/elections-integrity.html#data

@katestarbird
https://medium.com/@katestarbird/a-first-glimpse-through-the-data-window-onto-the-internet-research-agencys-twitter-operations-d4f0eea3f566

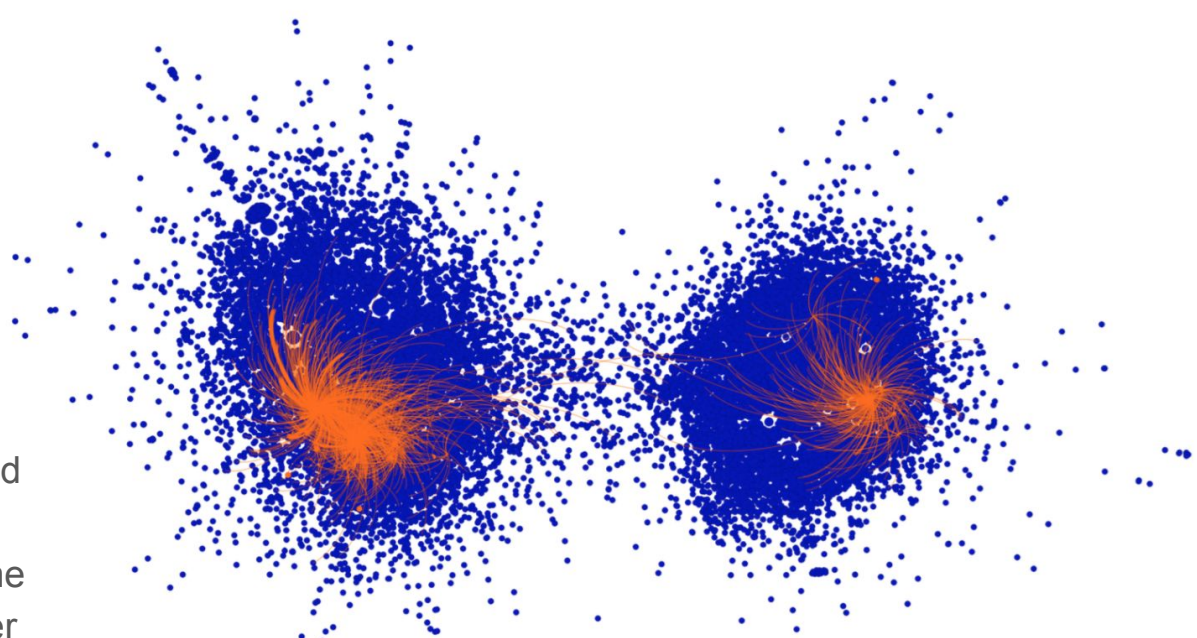


@katestarbird

https://medium.com/@katestarbird/a-first-glimpse-through-the-data-window-onto-the-internet-research-agencys-twitter-operations-d4f0eea3f566


Tweets and Retweets per Week
500000
400000
300000
200000
100000
0
2010-01-01
2011-01-01
2012-01-01
2013-01-01
2014-01-01
2015-01-01
2016-01-01
2017-01-01
2018-01-01
RTs of EN tweets
English tweets

Accounts that tend to retweet each other related to the #BlackLivesMatter Movement



https://medium.com/s/story/the-trolls-within-how-russian-information-operations-infiltrated-online-communities-691fb969b9e4

Ethical Concerns?

11-830: Computational Ethics for NLP

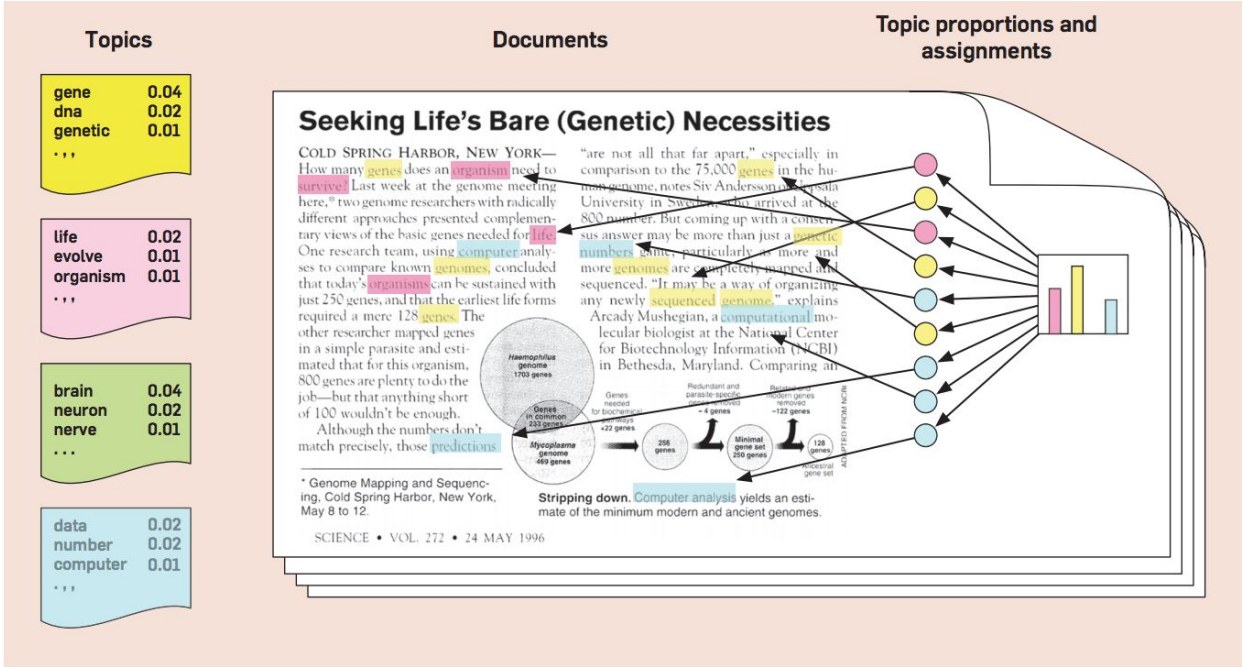
Methodology

Overview [Grimmer & Stewart, 2013]

- Classification
  - Hand-coding + supervised methods
  - Dictionary Methods
- Time series / frequency analysis
- Scaling (Map actors to ideological space)
  - Word scores
  - Word fish (generative approach)

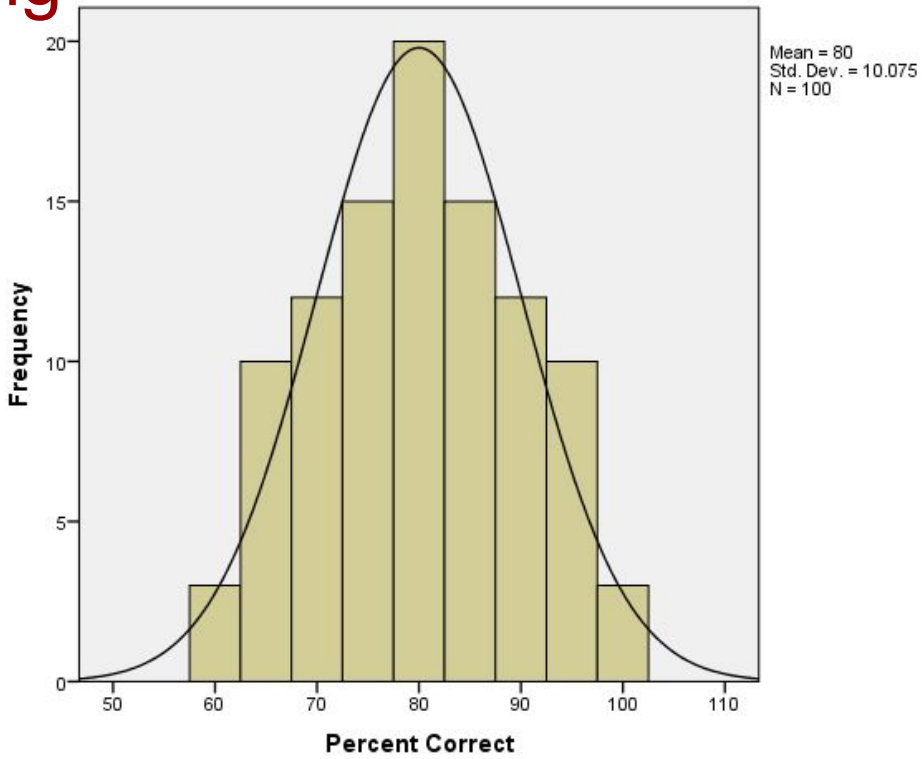
- Clustering (when classes are unknown)
  - Single-membership (ex. K-means)
  - Mixed membership models (ex. LDA)

Topic Modeling: Latent Dirichlet Allocation (LDA)



General Statistical Modeling

- Given some collection of data:
  - Assume you generated this data from some model
  - Estimate model parameters
- Example:
  - Assume you gathered data by sampling from a normal distribution
  - Estimate mean and stdev

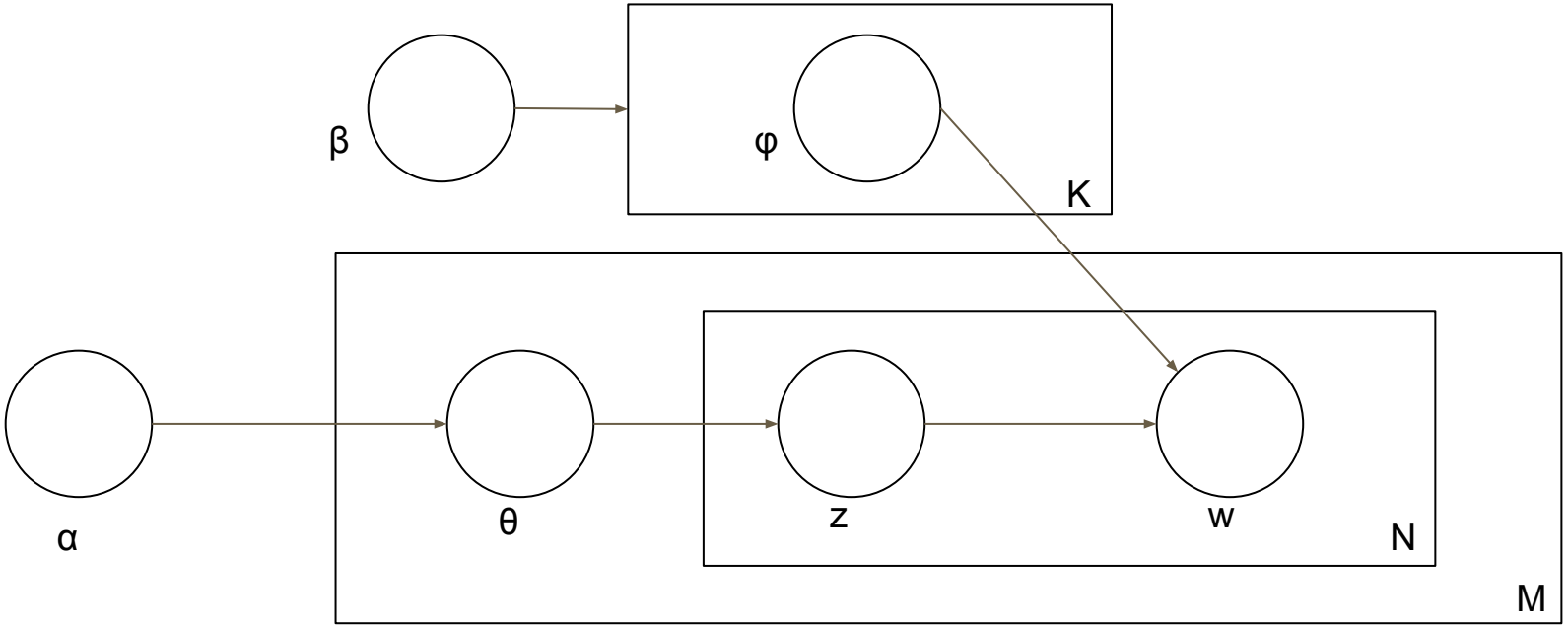


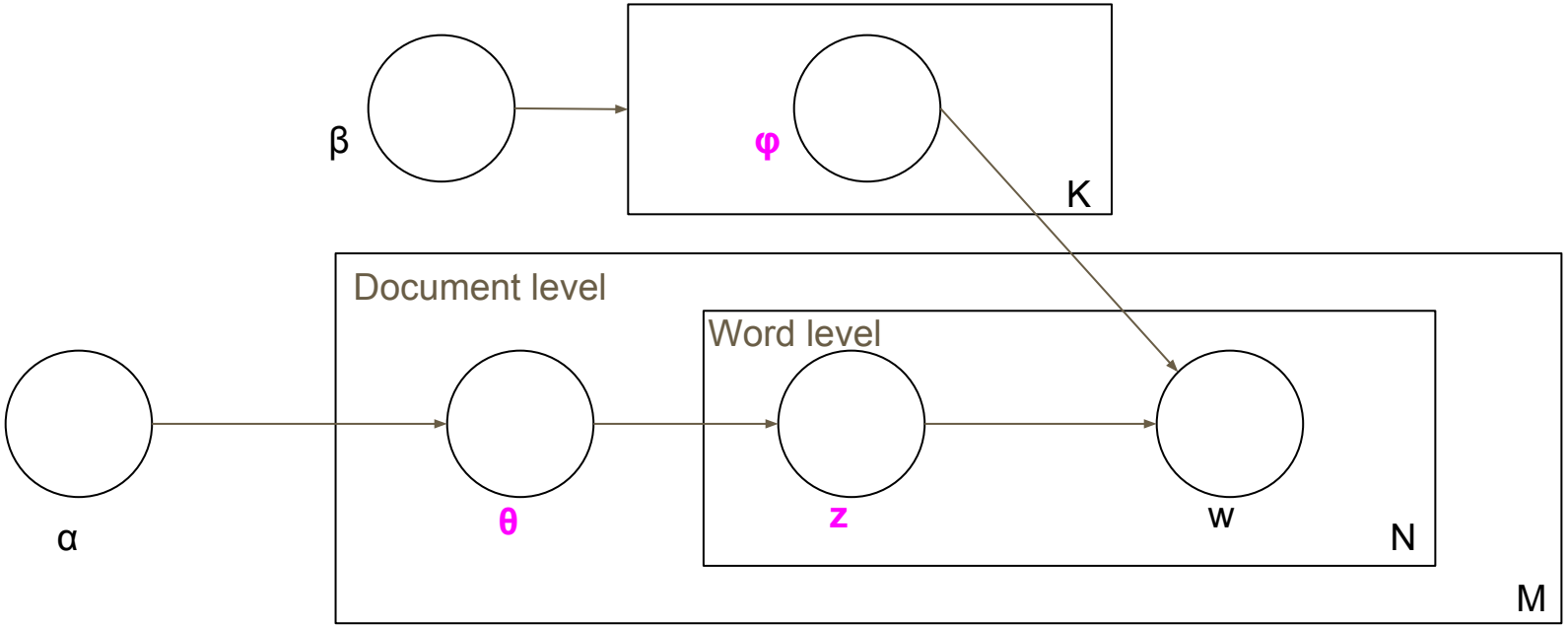
LDA: Generative Story

- For each topic k:
  - Draw $\varphi_k \sim Dir(\beta)$
- For each document D:
  - Draw $\theta_D \sim Dir(\alpha)$
  - For each word in D:
    - Draw topic assignment $z \sim Multinomial(\theta_D)$
    - Draw $w \sim Multinomial(\varphi_z)$

φ is a distribution over your vocabulary (1 for each topic)

θ is a distribution over topics (1 for each document)


β
φ
K
α
θ
z
w
N
M



θ, φ, z are latent variables
α, β are hyperparameters
K = number of topics; M = number of documents; N = number of words per document

Recap: General Estimators [Heinrich, 2005]

Goal: estimate θ, φ

$$p(\theta, \phi, z|w) = \frac{p(w|\theta, \phi, z)p(\theta, \phi, z)}{p(w)}$$

- MLE approach:
  - Maximize likelihood: **p(w | θ, φ, z)**
- MAP approach
  - Maximize posterior: **p(θ, φ, z | w)** OR **p(w | θ, φ, z) p(θ, φ, z)**
- Bayesian approach
  - Approximate posterior: p(θ, φ, z | w)
  - Take expectation of posterior to get point estimates

LDA: Bayesian Inference

Goal: estimate θ, φ

Bayesian approach: we estimate full posterior distribution

$$p(\theta, \phi, z|w) = \frac{p(\theta, \phi, z, w)}{p(w)}$$

p(w) is the probably of your data set occurring under *any* parameters -- this is intractable!

Solutions: Gibbs Sampling [Darlington 2011], Variational Inference

Sample Topics from NYT Corpus

#5	#6	#7	#8	#9	#10
10	0	he	court	had	sunday
30	tax	his	law	quarter	saturday
11	year	mr	case	points	friday
12	reports	said	federal	first	van
15	million	him	judge	second	weekend
13	credit	who	mr	year	gallery
14	taxes	had	lawyer	were	iowa
20	income	has	commission	last	duke
sept	included	when	legal	third	fair
16	500	not	lawyers	won	show

LDA: Evaluation

- Held out likelihood
  - Hold out some subset of your corpus
  - Says NOTHING about coherence of topics
- Intruder Detection Tasks [Chang et al. 2009]
  - Give annotators 5 words that are probable under topic A and 1 word that is probable under topic B
  - If topics are coherent, annotators should easily be able to identify the intruder

LDA: Advantages and Drawbacks

- When to use it
  - Initial investigation into unknown corpus
  - Concise description of corpus (dimensionality reduction)
  - [Features in downstream task]
- Limitations
  - Can’t apply to specific questions (completely unsupervised)
  - Simplified word representations
    - BOW model
    - Can’t take advantage of similar words (i.e. distributed representations)
  - Strict assumptions
    - Independence assumptions
    - Topics proportions are drawn from the same distribution for all documents

Beyond LDA

Problem 1: Topic Correlations

- LDA
  - In a vector drawn from a Dirichlet distribution (θ), elements are nearly independent

- Reality
  - A document about biology is more likely to also be about chemistry than skateboarding

Solution to Problem 1: Correlated Topic Model [Blei and Lafferty, 2006]

- For each topic k:
  - Draw $\varphi_k \sim Dir(\beta)$
- For each document D:
  - ~~Draw $\theta_D \sim Dir(\alpha)$~~ Draw $\eta_D \sim N(\mu, \Sigma)$; $\theta_D = f(\eta_D)$
  - For each word in D:
    - Draw topic assignment z ~ Multinomial(θ_D)
    - Draw w ~ Multinomial(φ_z)

Σ = Topic covariance matrix

φ is a distribution over your vocabulary (1 for each topic)

θ is a distribution over topics (1 for each document)

Solution to Problem 1: Correlated Topic Model [Blei and Lafferty, 2006]

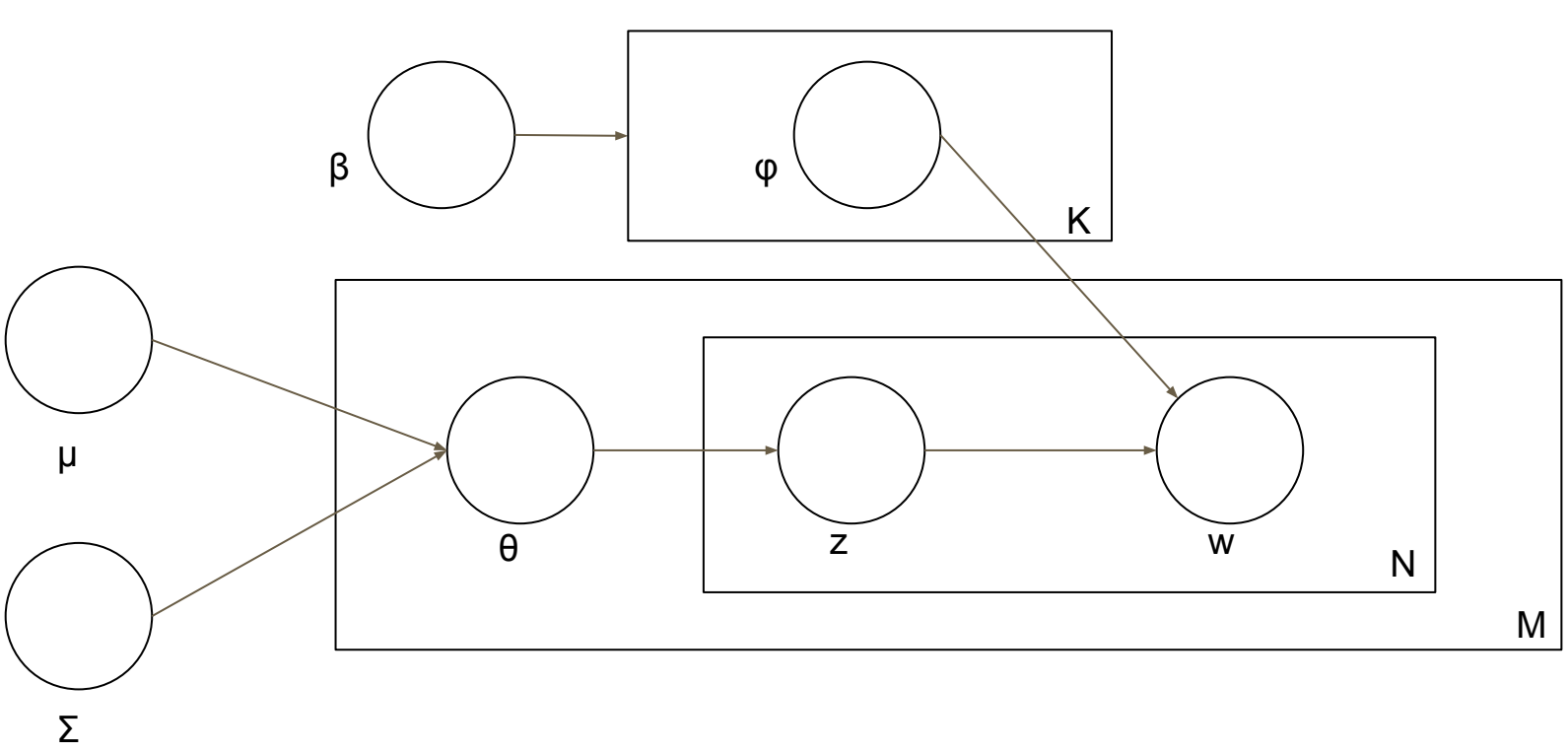
- For each topic k:
  - Draw $\varphi_k \sim Dir(\beta)$
- For each document D:
  - ~~Draw $\theta_D \sim Dir(\alpha)$~~ Draw $\eta_D \sim N(\mu, \Sigma)$; $\theta_D = f(\eta_D)$
  - For each word in D:
    - Draw topic assignment z ~ Multinomial(θ_D)
    - Draw w ~ Multinomial(φ_z)

Σ = Topic covariance matrix

Warning: Inference is harder!

φ is a distribution over your vocabulary (1 for each topic)

θ is a distribution over topics (1 for each document)


β
φ
K
μ
θ
z
w
N
M
Σ

Problem 2: Topics are drawn from same prior for all documents

- LDA
  - The topic distributions (θ) are drawn from the same distribution Dir(α) for all documents
- Reality
  - We often use LDA to look at how topics vary across documents
  - Example
    - We run LDA on a corpus of campaign speeches.
    - Look at topic prevalence in Republican speeches and Democratic speeches
    - Conclude Republicans talk about immigration more than Democrats
  - But we've assumed that all speeches are drawing topics the same way

Solution: Structured Topic Model [Roberts et al. 2016]

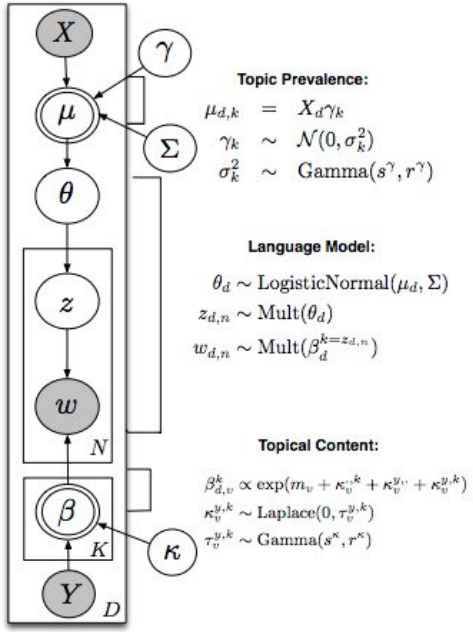


Figure 1: Plate Diagram for the Structural Topic Model

Topical prevalence: the proportion of document devoted to a given topic

Topical content: the rate of word use within a given topic

X - matrix of covariate information
Y - matrix of covariate information

Example:

- Analyze a corpus of news articles
- Topic prevalence covariates (X): date article was written, news agency
- Topic content (Y): news agency [do different agencies cover topics in different ways?]

Solution: Structured Topic Model [Roberts et al. 2016]

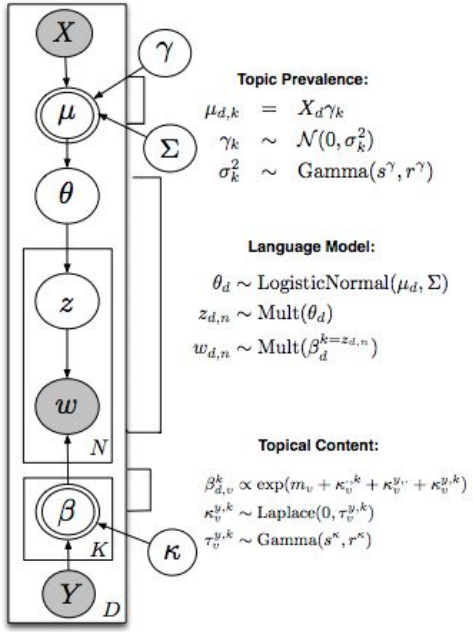


Figure 1: Plate Diagram for the Structural Topic Model

Topical prevalence: the proportion of document devoted to a given topic

Topical content: the rate of word use within a given topic

X - matrix of covariate information
Y - matrix of covariate information

Key contributions:

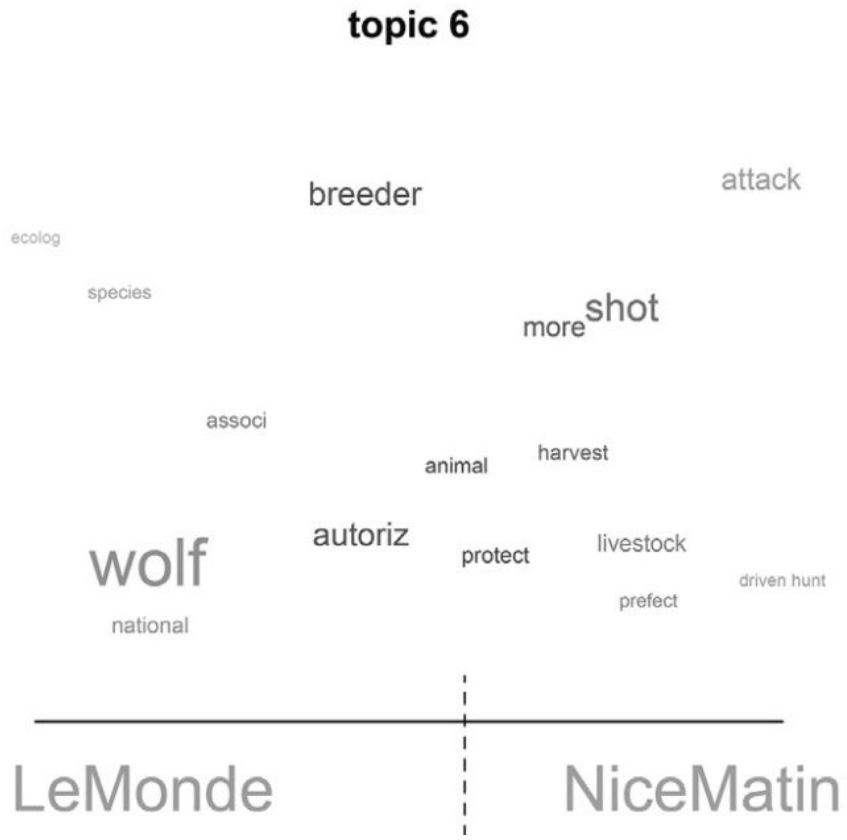
- Flexibly incorporate document-level metadata
- Allows correlations between topics

STM Example

21-year corpus on media coverage of grey wolf recovery in France

Nice-Matin = local newspaper
Le Monde = national newspaper

Topic 6: “Lethal Regulation”



https://www.structuraltopicmodel.com/
[Chandelier et al. 2018]

Summary

- Aspects of social science questions
  - Hard-to-define research questions
  - Messy data
  - “Explainability”
  - Ethics
- Topic Models
  - Generative story of LDA
  - LDA limitations and extensions

Why Computational Social Science?

“Despite all the hype, machine learning is not a be-all and end-all solution. We still need social scientists if we are going to use machine learning to study social phenomena in a responsible and ethical manner.” [Wallach 2018]

References

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*3.Jan (2003): 993-1022.
- Blei, David, and John Lafferty. "Correlated topic models." *Advances in neural information processing systems* 18 (2006): 147.
- Chandelier, Marie, et al. "Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling." *Biological Conservation* 220 (2018): 254-261.
- Chang, Jonathan, et al. "Reading tea leaves: How humans interpret topic models." *Advances in neural information processing systems*. 2009.
- Darling, William M. "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling." *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011.
- Gregor, Heinrich. "Parameter estimation for text analysis." *Technical report* (2005).
- Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21.3 (2013): 267-297.
- King, Gary, Jennifer Pan, and Margaret E. Roberts. "How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American Political Science Review* 111.3 (2017): 484-501.
- Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* 111.515 (2016): 988-1003.
- Roberts, Margaret E., et al. "The structural topic model and applied social science." *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. 2013.
- Wallach, Hanna. “Computational social science ≠ computer science + social data”. Commun. ACM 61, 3 (2018), 42-44. DOI: https://doi.org/10.1145/3132698